DOCUMENT RESUME

ED 295 990 TM 011 841

AUTHOR Mandeville, Garrett K.
TITLE An Evaluation of PET Using Extant Achievement Data.
PUB DATE Apr 88
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Elementary School Teachers; *Mathematics Achievement; *Reading Achievement; Teacher Education; *Teacher Effectiveness
IDENTIFIERS Hunter Instructional Model; *PET Training; South Carolina

ABSTRACT

Data gathered under various statewide testing programs in South Carolina were used to assess the effects of PET (data base) training for teachers on student achievement. Reading and mathematics achievement data for students in grades 1 through 4 in South Carolina were compared for teachers who had received PET training and those who did not. A total of 4,289 teachers' records were analyzed. The amount of training received by a typical PET trainee was small. Most of the PET teachers were in their first or second post-training year. There was too much residual variation in classroom achievement, as operationalized in the study, for the effects of PET training to be determined. Any differences existing at present are not educationally meaningful. It must be concluded that the Hunter Model has not been adequately evaluated. Three tables and one flowchart summarize the findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295990

An Evaluation of PET Using Extant Achievement Data

Garrett K. Mandeville

University of South Carolina

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GARRETT MANDEVILLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented as part of a symposium entitled
"South Carolina's Implementation of the Hunter Model:
Issues and Answers"
At the annual meeting of the American Educational Research Association
New Orleans, April 1988

BEST COPY AVAILABLE

TM011 841

In this paper I will show how test data, gathered under various statewide testing programs, were used to assess the effects of PET training upon student achievment. The paper will be presented in four sections.

In the first section, the details of the development of the data base will be described. This data set contains a large sample of 1st through 4th grade teachers in South Carolina and the aggregate reading and math performance of students they instructed in language arts and mathematics. Teachers who received PET training prior to the beginning of the 1986-87 school year (August 1986) and those who had not received PET training as of February 1987 were included; any teachers who were trained between August 1986 and February 1987 were excluded. These teachers all taught regular, self-contained classes. This requirement was necessary in order that the reading and math performance of their students in the Spring of 1987, adjusted for prior achievement, could be properly attributed to them. In order to insure that this attribution was correct, a number of restrictions were applied and various teachers were "lost." Except for these deleted teacher records, and those trained in PET between August 1986 and February 1987, the data base is statewide in nature.

The second section presents the results of analyzing these data to determine whether PET training has had a measureable effect on student achievement. In the third section, the subset of teachers from the larger data base who participated in the survey will be considered. The objective here will be to relate student achievement to variables and other information gathered via the survey. The final section will contain a summary and discussion of the results.

Building the Data Base

The reader will recall from Rivers' presentation that we began by matching the PET data base and the BEDS data base. This served to identify teachers of

self-contained classes and also located the school and grade-level in which the teacher was teaching. Upon incorporation of the information on lunch status, the samples of PET2, PET1, PET0, and NOPET teachers to be surveyed were drawn. At this time, of course, we could not be certain that student achievement could be linked to these records since the Spring 1987 testing had not taken place.

In developing the data base to assess student achievement as a function of PET training, a different data processing strategy was used. This was done to insure that the teacher-to-student achievement link was a correct one. The steps and rationale are described in detail below and in the flow chart in Figure 1.

Creation of BEDS Data File

The first step was to more completely utilize the information from the BEDS file. As before, only records of teachers in grades 1-4 who indicated that they were teaching in "regular", self-contained classrooms were retained. However, to guard against miscoded data further more stringent criteria were applied. In order to do so, the information on up to eight activity periods was considered. An activity period is simply a unit of time during which a teacher is engaged in some activity and teachers code the type of activity, number of children involved, duration, and a few other items for each one. Generally these are instructional activities but free periods for planning, etc., are also included.

To guard against situations in which large numbers of students were engaged in alternate activities which were not under the direction of their teacher, the difference between the maximum and minimum enrollments during these eight activity periods was computed and if this exceeded two, the record for the teacher was eliminated. The mean enrollment during these eight periods was also computed and the teacher record was deleted if fewer than 10 students,

on the average, were under his/her direction. Finally, since student achievement in reading and math were to be considered, it was important to insure that the teacher was responsible for instruction in these two areas. Thus, if the codes for language arts and math instruction did not both occur during the eight activity periods, the teacher was not considered further. Teacher age, years of teaching experience, race and sex were also obtained from the BEDS data base. Because of the conservative nature of the criteria applied, the 6049 teacher records which survived this scrutiny surely represented teachers of typical self-contained classrooms.

Creation of Aggregated Test Score Data File

For a number of years, all public school pupils at most grade levels in South Carolina have been tested each spring with either a criterion referenced test created by the state under the Basic Skills Assessment Program (BSAP) or the norm referenced Comprehensive Tests of Basic Skills (CTBS). Relevant to this study is the fact that the BSAP tests are used in grades 1-3 and the CTBS in grade 4. In addition, 1st graders are tested with the Comprehensive Skills Assessment Battery (CSAB) at the beginning of each year.

One of the ways these scores are used is in the determination of which schools are to receive funds distributed under the South Carolina School Incentive Reward Program (SIRP). Briefly, student records are matched year-to-year and statewide regression analyses are conducted in reading and math at each grade level. Students repeating the grade (except for the 1st) are matched in addition to nonrepeaters but, of course, separate regression equations are produced for each group. The residuals from these regressions, after standardization (they are "studentized" and multiplied by 10) and aggregation to the school level, serve as the basis for awards.

The matched data files were created by the author for the South Carolina

Department of Education (SCDOE) using a computer program written for that purpose. The algorithm, written in the FORTRAN 77 language, has been used effectively to produce the SIRP data for two years. (For more information on the mismatch and missed match rates, both of which are typically less than 1% prior to visual editing, and less than .5% thereafter, see Mandeville, 1988.)

In addition to the residuals, these data files contain a large number of other variables. Those relevant to this study are the student's eligibility for a subsidized lunch and the name from the so-called "Classroom Header", a record which precedes each group of student records. Usually the Classroom Header contains the name of the teacher.

The data sets for each of the four grades were then analyzed to obtain summary information for the matched student records associated with each Classroom Header. More specifically, aggregation across the set of matched observations following a header produced each of the following: 1) averages for the residual scores in reading and the residuals in math, 2) the percentage of students repeating the grade (except 1st grade), 3) the percentage of students eligible for free or reduced price lunches, and 4) the number of matched records with contained these data elements. During these analyses the original data files were accessed and the total number of posttest records associated with each header (whether matched or not) was also obtained. To be retained at this point, at least 10 records had to be included in the aggregation. The resulting data set contained 8491 records: 2551 1st grade; 2138 2nd grade; 2058 3nd grade; and 1744 4th grade.

The aggregated residuals on this data set will provide the evidence on the reading and math performance of each class. They reflect whether the typical student in that class performed at a higher (or lower) level on the posttest than was true of the typical student statewide with the same pretest score(s).

If more practice in applying the PET model (since initial training) does promote teaching skills and, therefore, student learning, we would expect the highest averages for the students of PET2 teachers, with the averages of PET1, PET0 and NOPET teachers following in sequence.

Merging the BEDS and Test Data Files

Next the BEDS and test data files created above were merged using a modification of the matching program mentioned earlier. The linking variable here was the teacher name. During this matching run, records for which the school codes and/or grade levels from the two sources did not agree were deleted. This was done for two reasons. First, it was felt that this would be a good way to identify (and delete) computer mismatches; second, for those teachers whose assignments did actually change between the fall and the spring, it seemed the prudent course to exclude them from the study. They might have, for example, been teaching their current class for only a few months in which case it would be inappropriate to impute its performance to their instruction. The number of records linked was 5608.

Merging the BEDS/Test Data File and the PET Teacher File

The final use of the matching algorithm was to link the records produced above with the file containing the names of the PET trained teachers. Unfortunately, besides the teacher's names and the date(s) they completed various aspects of the training, the only other information available was the code of the school district of employment. (School codes would have made it easier to check the validity of the matches but, of course, school assignments are not permanent). Nevertheless, the computer program linked 1606 records which, except for editing, represent the PET teachers in self-contained classes in grades 1-4 for whom achievement data for their classes were available. The 4002 nonmatched BEDS/DATA records, on the other hand, represent the teachers

with the same qualifications but who were not PET trained.

Incorporating Information from District Test Coordinators

In theory, the student records that follow the Classroom Header Record should represent the teacher's class but there might conceivably be some situations in which this was not true. For example, it was reported to the researcher that in the past some schools, for security or other reasons, have directed teachers to switch classes during testing. If the Classroom Headers contained the name of the test monitor rather than the classroom teacher, our imputations of student performance to teachers would obviously be incorrect.

In order to guard against this possibility, the Test Coordinator's of each of the 91 school districts in South Carolina were sent a computer printout of the schools in their district which contained any grade levels in the 1-4 span. They were apprised of our objective and requested to indicate whether our imputation scheme would be valid. Since the Test Coordinators are in charge of the details associated with the statewide testing programs in their respective districts, they were suggested by the SCDOE as the individuals most qualified to provide this information. Although follow-up telephone calls were necessary in a few cases, complete information was eventually obtained. When these further restrictions were applied to the two data sets, the number of PET teacher records was reduced from 1606 to 1305 and the number of NOPET records was lowered from 4002 to 3483.

Visual Editing of the Computer Matches

The final step in the process was to assess the veracity of the computer matches. The BEDS/DATA matches were the least likely to be in error because, in addition to the linkages based on the similarity of the names, the two records designated the same school and grade. For the PET trained teachers, however, the third match--based only on the teacher name and school district--

was more likely to be in error. In order to identify possible computer mismatches, a rather conservative set of rules was developed and applied during visual editing of each two-way and three-way matched record. (See the Appendix for these rules.) When they were applied, 26 PET and 6 NOPET records were deleted leaving 1279 PET and 3477 NOPET records.

Editing on Class Size Differential

During the computer runs to compute the results, one final edit was applied to all records. This involved a comparison of the mean enrollment over the first eight activity periods as reported by the teacher on the fall BEDS form and the number of students associated with the matched Classroom Header. Because of student mobility and/or illness during testing it would be quite unusual if these two numbers were identical but too large a discrepancy could signal an as yet undetected mismatch. What was more likely was that the class either gained or lost a relatively large number of students over the course of the year but, if this was the case, its removal as atypical seemed prudent. The criterion selected was that if the two estimates of class size differed by more than five children the record would be deleted. Although arbitrary, this number seemed appropriately conservative. A total of 1241 PET and 3378 NOPET records satisfied this criterion.

Also during data processing, the PET records were classified according to the amount of time between completion of the (first cycle of) training and the beginning of the 1986-87 school year. The categories were as follows:

PET2 - Completed cycle 1 training by August, 1984 (Two full years prior to the beginning of the 1986-87 school year)

PET1 - Completed between Sept. 1984 and August, 1985 (One full year prior to the beginning of the 1986-87 school year)

PET0 - Completed between Sept. 1985 and August, 1986 (Less than one full

year prior to the beginning of the 1986-87 school year)

OTHER - Completed between Sept. 1986 and February 1987

The OTHER teacher group, who were being trained while they were teaching the classes from which the achievement data were to be derived, were eliminated from the study (n = 330). The number of PET teachers which survived these many edits, by category, were as follows: PET2 - 26; PET1 - 248; and PET0 - 637. Because of the small number of PET2 teachers relative to the sizes of the other groups, they were combined with the PET1 group for purposes of analysis. (For some of the later analyses involving the survey data, the original four groups were retained.) Thus, the PET12 group completed their training at least one year prior to the beginning of the 1986-87 school year. Table 1 reflects the final sample size by grade and category (for convenience this independent variable will be denoted as CAT).

Table 1 about here

Results Based on All Matched Teachers

Description and Comparison of the Three Samples

For the most part, the three groups were similar on the demographic variables available. Over 98% of the teachers were female and this percentage was consistent across the three groups for each grade. Nor did the three groups differ in age or the number of years of experience. Although a two-way analysis was not performed, there was a tendency for slightly older and more experienced teachers to be assigned to the higher grades. For example, the average ages of teachers by grade were: 1st grade-36.6 years; 2nd grade-37.1 years; 3rd grade-38.1 years; and 4th grade-38.7 years. Similarly, the average number of years of teaching experience ranged from 10.9 for 1st grade teachers

to 12.3 for 4th grade teachers.

There were, however, significant differences in teacher race as a function of CAT in grades 1 (Chi Square=7.34, df=2, p=.025) and 2 (Chi Square=7.18, df=2, p=.028). At grade 1 the data suggest a smaller percentage (76.1) of white teachers in the PET0 group than the other two groups (PET12, 84.6% and NOPET, 83.3%) but post hoc tests using a procedure outlined by Marascuilo and McSweeney (1977, p. 141) failed to judge any pairwise differences as significant at the .05 level. At grade 2 the PET12 percent white (88.0) was statistically larger than the corresponding PET0 value (72.8%) but not the NOPET value of 78.5%. Although equivocal, the data suggest a bias in favor of larger proportions of white 1st and 2nd grade teachers participating during the first two years of PET training and a correspondingly lower proportion during the third year of implementation.

Information on variables describing the classes was also considered. The total number of Spring 1987 test records was selected as the best estimate of class size and compared for the three groups for all grades. One significant difference occurred at grade 2 (F=3.11, df=2, P=.045) but post-hoc tests using Scheffe's Method (because of the vastly divergent sample sizes) provided no significant pairwise differences. The largest mean difference indicated that PET0 classes were about one student larger on the average than PET12 classes.

When the percentage of children eligible for free or reduced price lunches was considered, significant differences in the means were found for grades 1 and 2. For grade 1 (F=6.77, df=2, P<.001) post-hoc tests indicated that the classes of the PET12 and NOPET groups contained significantly smaller average percentages, on the average, of children in this low SES condition, than the PET0 classes. The means were 44.2% and 49.5% for the PET12 and NOPET classes, respectively, and 55.5% for the PET0 classes. The means of the PET12 and NOPET

groups were not significantly different. Although the general trend in the data was similar in the 2nd grade, here the mean for the PET12 group (39.0%) was significantly lower than the means for the NOPET (49.0%) and PET0 (50.0%) groups with the latter two groups obviously not different.

Another datum concerning the classes which was available was the scheduling during the day of language arts and math instruction. (This had been collected to assure that the teachers were responsible for this instruction since the achievement measures were reading and math tests.) The dominant theme in the state is to teach language arts during the first period and math during the second with over 82% of the sample doing so. (The only other patterns which are at all frequent generally also treat these two subjects early in the day, at least before the end of the 4th activity period which probably corresponds to lunch in most cases.) In each grade, over 90% of the PET12 teachers used this "L-1,M-2" approach to language arts and math instruction at each grade level whereas this was only true for the other two groups at the 4th grade. The only difference which was statistically reliable (at the .05 level) was at grade 3 (Chi Square=10.7, df=2, P=.005) where the percentages were as follows: PET12-95%; PET0-88%; and NOPET-80% but the differences were in the same direction and of a similar magnitude in grades 1 and 2 and the associated P-values were between .05 and .10.

The final variable was the percentage of students repeating the grade which applied to grades 2-4 only. Because about 80 percent of the classes had 5% or fewer repeaters, this variable was collapsed into four categories and Chi Square was used for the comparisons. No differences were detected among the three groups at any grade level.

Comparison of the Three Samples on Achievement

The fact that at some grade levels the three groups differ, or at least

appear to differ, with respect to the SES composition of the classes, suggests a need to control this variable. The problem of controlling for selection bias is not a trivial one (see, e.g., Cook & Campbell, 1979). Fortunately, the reliability of the information on the percentage of students eligible for free or reduced price lunches (LUNCHPCT) should be high so it is an appealing choice as a covariate. In working with South Carolina test data, when aggregated to the school level, it has always proven to be a powerful predictor of achievement. It was decided that initially both an ANOVA (which would be known to be biased) and and ANCOVA (possibly also biased) would be used to analyze these data. Later, if the results warranted it, other approaches to dealing with the selection bias would be considered.

Preliminary to the ANCOVA tests, the assumption of homogeneity of the regression coefficients was tested. Using the .10 level to increase the power of the test, the only case where the assumption was not satisfied was for 2nd grade reading ($F[2,1126]=3.1$, $P=.045$). In each case, LUNCHPCT was a significant ($P<.001$) predictor of the dependent variable and, as would be expect, all slopes were negative.

None of the F-tests from the ANOVAs or ANCOVAs was significant at traditional levels of significance. For purposes of discussion, the unadjusted means and P-values from the ANOVAs and, where appropriate, the corresponding statistics from the ANCOVAs are presented in Table 2 below.

Table 2 about here

We note that adjustments for LUNCHPCT are generally small for the two PET groups and virtually nonexistent for the much larger NOPET sample. As expected, the adjustments tend to decrease the averages slightly for the

apparently advantaged PET12 group and cause small increases for the PET0 group. The reasons that the null hypotheses were not rejected is also quite clear when the sizes of the mean differences are considered in light of the residual standard deviations which are included in the right hand column of the table. Means for the classes in each of the three samples typically differ from the mean by 4 to 5 units. Thus, even with samples as large as these, the small mean differences are well within sampling variation. We must conclude that no systematic effects associated with PET training have been manifest.

Results Based on the Subsample of Teachers Surveyed

Although the above results were hardly ambiguous, it was deemed advisable to consider the achievement data for the subsample of teachers who were involved in the survey. Considering those who returned useable survey data, the sample size of the survivors was reduced to 171. The smallest group was the PET2 teachers with n=24; the other three sample sizes were each about twice as large with ns of 54, 47, 46 for the PET1, PET0 and NOPET groups respectively. Because the sample sizes were reasonably comparable, PET2 and PET1 teachers were not combined for these analyses. The four groups did not significantly differ on sex, race or age. For years of teaching experience ($F[3,166]=2.59$, $P=.054$) the data suggested that PET0 teachers were less experienced (M=9.2 years) than the other three groups for which the means exceeded 12 years.

Achievement

The data analysis should reflect the fact that LCHCAT was built into the sampling design. Since the sample sizes were too small to conduct two-way analyses at each grade level, the data were collapsed over grade and a two-factor (4x4) ANOVA with free/reduced price lunch category (LCHCAT) and CAT as the factors was conducted. As could be predicted from the earlier findings,

the results for neither reading nor math approached significance.

Relationships Between Achievement and Survey Variables

At this point the data set was delimited to PET trained teachers only and a number of analyses seeking to assess the relationship between a number of survey related variables and achievement were conducted. One approach to assessing the relationships between many of the survey related variables and the achievement measures would have been to correlate them. At least this would seem reasonable for those survey variables which could be considered to be measurements on at least a "quasi interval" scale. The distributions of many of these variables were highly skewed, however, reflecting the teachers general tendency to respond in a very positive fashion to the questionnaire. In these situations the problem of the effect of the marginal distributions on the magnitude of correlations suggested that a different approach be found.

Various ANOVA strategies could be considered. For example, if a survey variable, X say, were subdivided into categories, a one-way ANOVA with X as the independent variable and the achievement measures as the dependent variables would shed some light on whether X was related to student achievement. Since the length of time since PET training was a primary focus of the study, however, two-way ANOVAs with CAT and X as the independent variables seemed more appropriate and this was the general approach selected. Table 3 provides a summary of these analyses. Included is a description of how the responses to X were grouped, an indication of the statistical significance for reading and math for the X main effects or the CAT*X interactions. (The CAT main effects were never significant.)

Table 3 about here

Of the six general questions, none were significant at the .05 level but FPLANLES did achieve the 10% level for reading. This tentative "result" suggests that students of teachers responding that they used PET concepts daily in planning their lessons (M=.85) outperformed the students of those who indicated less frequent usage (M=-1.36), a result which makes sense.

The amount of coaching per year by either the PET trainer or the principal (taken as separate variables) was not related to student achievement nor was the simple coached/not coached variable. The nonsignificant results for TRAINER and PRINPAL suggest that quantitatively more training by either a PET trainer or the principal did not effect the residualized class means. Even more surprisingly, the pupils of teachers who reported no coaching did not achieve differently than those of teachers who reported at least some "official" coaching.

However, again at tentative significance levels, an interaction between the amount of "official coaching" (trainer or principal) did interact with CAT for both reading and math achievement. Inspection of the means did not suggest any meaningful interpretation of this result; since some of the cell sizes are quite small (they vary from 2 to 19) it is quite likely that chance variation is the most plausible explanation for this "almost significant" finding.

The results for the four scales assessing the perceived adequacy of the coaching produced only one significant result even at the less stringent .10 level. This tentative result occurred for math on SCALE4, the equivocation factor. The different bases of the F test and Tukey's Method became evident, as one of the pairwise differences in the means was significant at the .05 level using the T method. The only mean which was positive was for the group coded as "1" which had an average response in the 1 to 1.5 range. Since SCALE4 contained only two items, both reversed, this means that a "1.5" respondent

either originally marked "SD" to both items or SD to one and Disagree to the other. The group "1" mean (M=3.15) was statistically larger than the M=-2.14 for group "2"; the group "3" and "4" means were -.36 and -.64, respectively. These results are confusing but suggest, at least, that strong agreement on this scale might be related to better performance on the part of the teacher's students. Finally no differences were found as a function of the teacher's motivation for PET training.

Other Questions Related to or Suggested by the Survey

Based on the presumed importance of coaching to PET, it seemed reasonable to categorize PET teachers on the quantity of coaching reported and its perceived quality. The "quantity" portion was based on whether the teacher had received at least two official coaching observations per year since PET training (High) vs. less than two (Low). (High as used here should be considered relative; it is likely that PET proponents would suggest that more than two annual coaching observations are necessary.) For the quality portion two different definitions were used. One was based on the analysis scale (SCALE2) because it is the most sensible from a theoretical standpoint; the second was based on the equivocation scale (SCALE4) since the earlier results suggested that it might be related to achievement differences. For both scales, "High" quality was defined as score of 2 or less. This represents a typical response of agree or stronger. Subjects not classified as High were coded as Low quality.

The above set-up leads to four categories for each variable (they will be denoted QQ1 and QQ2 where QQ denotes quantity-quality). It was heartening that over 50% of the PET teachers were "HH" on QQ1 and that fewer than 10% were in the "LL" category. For QQ2, however, the percentages in the two extreme categories were about the same: "HH" - 21.7% and "LL" - 19.6%. None of the

results of the two-way ANOVAs approached significance.

Another way of utilizing this categorization scheme would be to select PET teachers who were in the "HH" category to represent those for whom the PET training and subsequent coaching approached might be considered "adequate." The performance of these teachers' pupils could then be compared to the performance of students of NOPET teachers. It is important to note, however, that if the results should happen to favor the adequately coached PET trainees, this would speak to an evaluation of PET as we would like to see it implemented, rather than as it is actually being implemented. Even with this qualification, the ANOVAs comparing the four groups were never significant, even at the less stringent .10 level.

For another analysis, the performance of teachers (actually their students) was compared as a function of who served as their coach. Those with one or more observation per year by the principal and fewer than one observation per year by a PET trainer were categorized as "P" (principal); those with at least one observation per year from each of these two sources were classified as "B" (both). (There were too few teachers who were observed by a PET trainer only to create a "T" category.) The comparison of these two groups with a t-test was nonsignificant for reading but significant for math ($t=2.15$, $df=93$, $P=.03$). The mean for the B group was 1.06 vs. a mean of -1.12 for the P group. The data suggest the efficacy of involving a PET trainer in at least a portion of the coaching observations.

An analysis suggested by the written comments to the survey was to delimit the analysis to fairly inexperienced teachers. (The comments were to the effect that "I wish I had known about PET when I began teaching.") Thus, this analysis was limited to teachers with 4 or fewer years of experience. There were only 28 teachers with this little experience and none were in the PET2

group. One-way ANOVAs were conducted with the three remaining levels of CAT serving as the independent variable. The result for reading, although not significant at the .05 level, was suggestive ($F[2,25]=2.92$, $P=.07$). The means were PET12, 2.63 (n=7); PET0, -1.25 (n=12); and NOPET, 1.83 (n=9). The corresponding ANCOVA with LUNCHPCT as the covariate did achieve significance ($F[2,24]=3.73$, $P=.04$) with the adjusted means comparable to the unadjusted means given above. Individual comparisons suggested that the reading average for PET12 teachers was larger than the corresponding average for PET0 teachers. No differences were found in math.

Since information on years of experience was derived from the BEDS file, i.e., its availability was not dependent of whether the teacher returned the survey, for completeness analyses paralleling the above were also conducted for the full data set. None of the results approached significance, however.

Discussion

One of the reasons for embarking on this research was to bring data to bear on the Hunter contention concerning the necessity of two years of adequate coaching for PET trained teachers to become skilled in the application of the model. From the perspective of this paper, this was taken to mean that their students would outperform similar students of untrained teachers on achievement tests. Unfortunately, few teachers were trained during the first year and our PET2 sample was small, probably too small to provide a definitive answer to this question. The fact that the students of the few PET2 teachers who received two or more "high quality" coaching observations per year scored no higher than the students of other teachers is suggestive. It is quite possible, however, that neither the quantity nor the quality would meet with the approval of Hunter or the proponents of her model.

As regards the more general question of how students of PET trained

teachers performed in comparison to students of NOPET teachers, the data appear to be sufficient to make a summary statement. There is too much residual variation in classroom achievement, as operationalized here, for whatever modest PET effects are operating to stand out. Thus it must be concluded that whatever differences currently exist, they are not educationally meaningful. I will attempt to justify this position in the paragraphs to follow.

Acceptance of the Null Hypothesis

Serious researchers will take issue with my rather strong inference--that current PET effects are not educationally meaningful--based on their knowledge that failure to reject the null hypothesis does not prove that it is true. That is to say, researchers prefer to reject null hypotheses because then the chance of making an incorrect inference is controlled at a relatively small probability, i.e., the Type I error rate. Attempts to place interpretations on statistically nonsignificant findings require an analysis concerning possible Type II error rates and the power of the statistical procedure against specified alternative hypotheses.

For the purpose of argument, let us say that a reasonable expectation is that in the first year after training (PET0), a PET trained teacher might foster a .1 SD increase in the performance of his/her typical student vs. how that student would have scored if the teacher had not been trained. (In the metric employed here, this is a 1 point advantage.) Let's add another .1 SD for those teachers who had a full year or more of practice (PET12). For those who argue that it is too much to expect an educational innovation to foster a one-tenth or two-tenth SD increase in student performance, a rule-of-thumb with which this writer is familiar requires a one-third SD effect for it to be considered educationally significant.

Armed with these "parameters", the ns from Table 1, and assuming a

residual variance of 25 (see Table 2), it is a straightforward calculation to estimate the probability of obtaining a statistically significant result (see, e.g., Winer, 1971, p. 220). The computations lead to power estimates which range from about .95 for grade 1 to .70 for grade 4 for .05 level tests. Most researchers would be exceedingly happy to employ statistical tests with power in this range.

Issues Related to the Use of Student Achievement

Some will argue that in evaluating educational innovations such as PET it is inappropriate to concentrate on student achievement as the only dependent variable and I concur with them. Others may argue that student achievement should not be considered at all; I disagree with them. Considering the amount of time and resources which are spent in measuring student achievement it is obviously viewed as an important, probably the most important, educational outcome. It does not seem unfair to expect educational innovations such as PET to effect achievement in some measureable way.

Of course, it is important that the tests used are valid for the purposes of measuring student learning. Fortunately, curricular validity is reasonably well assured for the BSAP tests since they were constructed to South Carolina specifications and the 11 reading and math objectives assessed have been articulated to teachers and their administrators for many years. The curricular validity of the CTBS is not as clear but, presumably, it was selected by SCDOE officials because it was reasonably consistent with state educational objectives.

Given valid tests, the researcher also has choices in the way to operationalize "classroom achievement." The approach used here has much to recommend it. For example, the use of the matched student records makes it reasonably certain that each datum represents a student who was under the

continuous instruction of the teacher throughout the year. Furthermore, each student's end-of-year achievement is adjusted, via the regression analysis, for earlier performance. The use of the SIRP residuals also has the advantage that students who are repeating the grade may be included in the analysis. It is difficult to conceive of an alternative approach which has all three of these advantages.

Selection Bias and External Validity

The data strongly suggest that selection factors operated, at least in grades 1 and 2, so that the first cohort of PET trained teachers (PET2) taught classes with fewer low SES students than was typical statewide. This selection bias, however, should operate to the advantage of these early PET trainees as regards the unadjusted results of the ANOVAs. Since no systematic differences were found favoring the PET2 teacher group, it is difficult to believe that their students performed in any systematically different way on the achievement tests than NOPET teachers.

The fact that a large number of cases were lost during the creation of the data base is an unhappy reality associated with this type of research. There is no reason to believe that teachers who were eliminated from the sample at the various decision points are any different from those who survived, however. Of course, the results do not pertain to teachers in departmentalized or team teaching environments or those who teach at higher grade levels.

Has the Hunter Model Been Fairly Evaluated?

The conclusion above of no meaningful educational effect on student learning does not necessarily reflect on the Hunter Model; rather, it reflects on the situation three years after the beginning of the PET implementation in South Carolina. As noted above, Hunter emphasizes the importance of regular formative evaluation by skilled coaches and anticipates that two years of

practice will be necessary for a teacher to internalize the principles and achieve an artistic level of teaching.

Based on our survey, the amount of coaching received by a typical PET trainee has not been large. Although we attempted to indirectly assess the quality of the coaching, the results for our four scales are not unambiguous. Although perceptions regarding three of the scales were quite positive, the results for Scale 4, our so-called equivocation factor, were mixed. This suggests that a number of coaches are approaching their task more from a summative than a formative evaluation framework. It is likely that the majority of these coaches are principals rather than PET trainers which would be consistent with the tentative finding that being coached by both principals and trainers was positively related to student achievement.

It is also true that a large percentage of the teachers in our PET sample were in their first or second post-training year. Therefore, most have not had time, according to Hunter, to achieve the artistic level she prescribes. Recent results from the SIM Project indicate that "...the Hunter Model works--it also takes a long time..." (Manatt, 1988). Since the SIM Project focused solely on student achievement as the dependent variable, there is reason to be hopeful. However, SIM used a total-systems approach which had a strong emphasis on improving administrators' skills in leading teachers.

With these facts in mind, it must be concluded that the Hunter Model has not been adequately evaluated. This study should, however, raise some warning signals concerning the South Carolina implementation. As is true of all educational innovations, PET is no quick fix. Nor is it an easy fix. The time problem will be solved with the quick passage of a few years. The coaching problem will need more careful attention.

Recommendations for Future Research

Given the rather negative overall results, it was hoped that the analyses relating student achievement to variables derived from the survey would provide a basis for recommendations regarding future PET training and/or subsequent coaching. In fact, based on earlier analyses of these data--prior to the implementation of many of the edits which were included to insure the efficacy of the data attribution scheme--a few tentative relationships between the quality of coaching and student achievement were identified. Most of these earlier findings were not supported when the sample was further delimited.

One which receive somewhat tentative support, however, may be noteworthy. I am referring to the act that students of PET teachers with Scale 4 (equivocation) scores which indicated that their coaching was most in harmony with the PET model, outscored students in the next group in math. Obviously, this result would have been less ambiguous if other post hoc comparisons had also been significant, but this was not the case. Although the other two group means were slightly negative, the sample sizes were relatively small, ranging from 16 to 35, so that the power of the statistical test was not large. Nevertheless, at the risk of overinterpreting this result, the following explanation is posited.

Coaching consistent with low (after item reflection) scores on the equivocation scale may be crucial for teachers to internalize the model. Coaching behaviors such as pointing out mistakes and rendering unsolicited opinions may be extremely detrimental for at least two reasons. First, they may thwart the type of coach-teacher interpersonal relationship necessary for coaching to function at all. Second, and more specifically, such behaviors probably discourage the type of reflective behavior on the part of the teacher which is necessary for him/her to progress independent of the coach.

The fact that the above finding was manifest in math and not reading

(although the means tended in the same direction) is also interesting. There are two, possibly interrelated explanations for this. The first is that math, being sequential in nature, is probably more amenable to direct instruction than reading. Second, work on school effectiveness indices (SEIs) by Mandeville and Anderson (1987) (and others) has shown that math achievement is more stable than reading. Their suggested explanation of this phenomena was that "young children are more likely to gain knowledge and skills in areas such as reading from sources outside the school than is true for areas such as mathematics." (Mandeville & Anderson, 1987, p. 213). Thus, it is reasonable to expect that PET and other similar programs may have more potency for affecting math than reading achievement.

Admittedly, the above discussion, related to a rather minor statistically significant result, must be considered to be very tentative. However, it does suggest that future research focused on the types of coaching behaviors assessed by the equivocation factor may be productive.

In light of that fact that our PET2 samples were quite small, the most obvious research study to suggest would be to wait a year or two and conduct a similar study. This may be impractical, however, since the pool of untrained teachers is diminishing rapidly. If no control group is available, a longitudinal design might be employed. This would allow for a profile of changes (hopefully, improvement) in student achievement over time and would clarify the pattern of whatever changes are manifest. For example, the power analysis presented earlier assumed a linear effect over time whereas Hunter's ideas suggest discontinuous improvement.

Another possibility would be to consider different operationalizations of classroom achievement. The mean, although exhibiting many nice statistical properties, does not allow the consideration of whether subgroups of students

perform differentially. Slopes-as-outcomes measures could also be considered.

Although not directly related to PET, a potentially fruitful line of research would be to conduct studies which attempt to further our understanding of the large variation between classes, e.g., adjusted classrooms SDs of about one-half a student SD. Of major importance would be to determine whether classroom achievement measures are stable year-to-year. Research on the cross-year stability of SEIs has yielded modestly high correlations (rs in the .40 to .60 range) for grade-specific SEIs (Mandeville, 1987). This finding indicates that different children, taught to a large extent by the same teachers, perform reasonably similarly.

References

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

Manatt, R. P. (1988). Personal communication to Janelle Rivers-Laney.

Mandeville, G. K. (1987). The stability of school effectiveness indices across years. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington DC.

Mandeville, G. K. (1988). A record linking system for research on school improvement. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Mandeville, G. K. & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. Journal of Educational Measurement, 24, 203-216.

Marascuilo, L. A. & McSweeney, M. (1977). Nonparametric and Distribution-Free Methods for the Social Sciences. Monterey, CA: Brooks/Cole.

Winer, B. J. (1971). Statistical Principles in Experimental Design (2nd ed.). New York: McGraw-Hill.

Table 1

Final Sample Size by Training Category and Grade

	Training Category (CAT)			
Grade	PET12	PET0	NOPET	Total
1	92	231	1233	1556
2	75	169	888	1132
3	58	138	727	923
4	49	99	530	678
Total	274	637	3378	4289

Table 2
Unadjusted Means from the ANOVAs
and Adjusted Means from the ANCOVAs
For Reading and Math Residuals
by CAT and GRADE

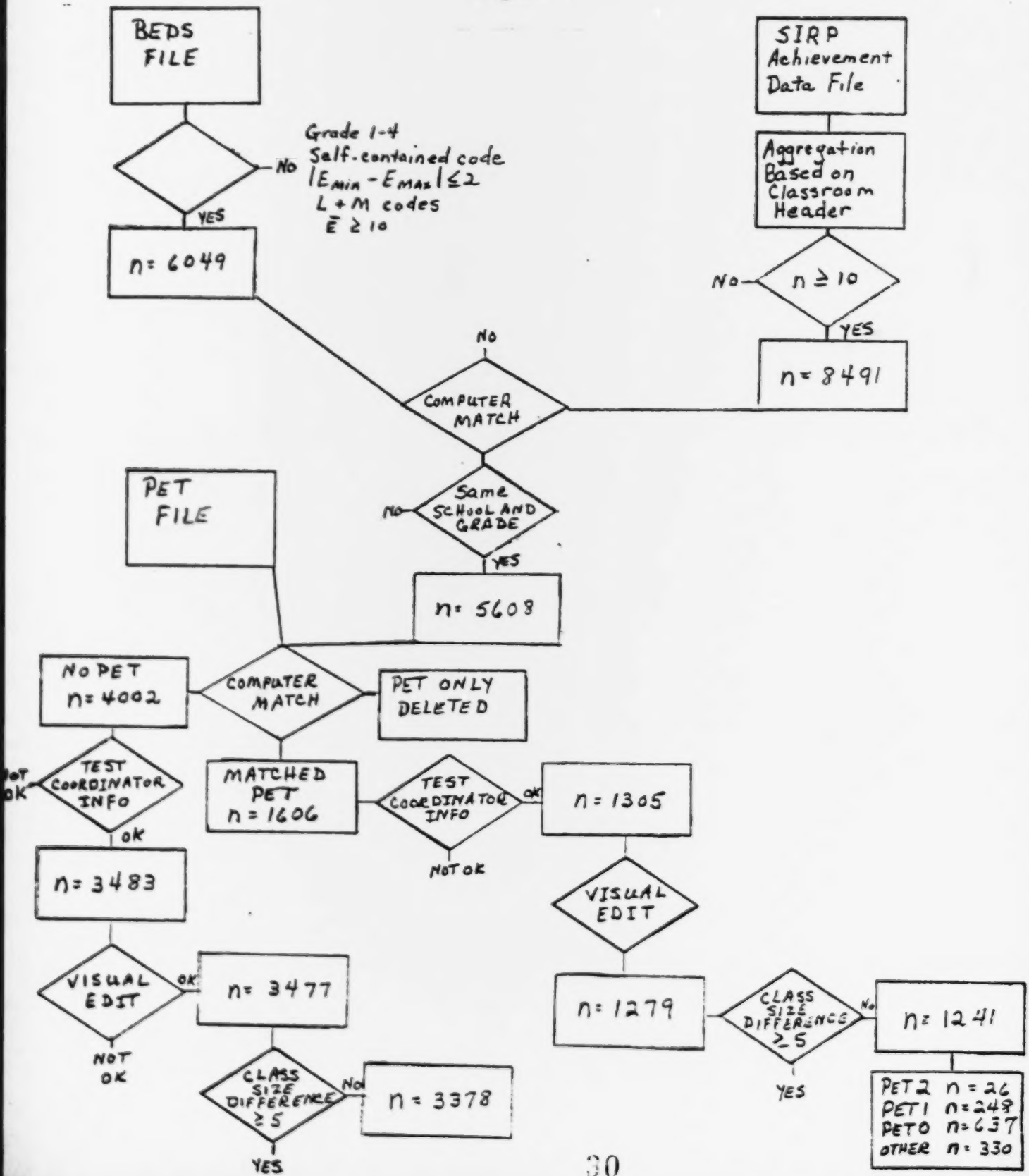
		Training Category (CAT)								
Grade	Subject	PET12		PET0		NOPET		P-values		ANC
		U	A	U	A	U	A	ANO	ANC	SD
1	R	.11	-.18	.82	1.10	.62	.59	.56	.12	5.1
1	M	-.06	-.22	.04	.18	.32	.30	.58	.57	4.8
2	R	.66	NA	-.08	NA	.48	NA	.29	NA	NA
2	M	.12	-.34	-.14	-.06	.71	.73	.13	.055	5.1
3	R	-.24	-.24	.98	.95	.14	.14	.08	.08	4.3
3	M	.08	.08	.73	.70	.38	.38	.71	.74	5.5
4	R	-.24	-.38	.19	-.38	.12	.09	.83	.51	3.8
4	M	.39	.32	.05	.14	-.09	-.08	.79	.80	4.8

Note: U and A above refer to "unadjusted" and "adjusted" means, respectively; ANO and ANC are used to refer to the ANOVA and ANCOVA P-values; and the "ANC SD" is the square root of the adjusted error mean square from the ANCOVA.

Table 3
Summary of Two-Factor (CAT by X) ANOVAs
Where X Represents Variables Derived from PET Teacher Survey
With Reading and Math Achievement as the Dependent Variables

Area Var X(Description)	# levels/Description	Info. on Fs Related to X and X*CAT Int. READ	 MATH
General Questions			
FCONTERM (Use PET concepts/terms)	4/1=daily, 3=2-3/wk, 4=weekly, 6=month, rare	NS	NS
FPLANLES (Use PET lesson plans)	2/1=daily, 2=less often	F=3.75 (P=.06)	NS
POSEFF (Pos effect on class perf)	3/1=SA, 2=Ag, 3=No op, Dis	NS	NS
INCRCONF (Incr confid tchg abil)	3/1=SA, 2=Ag, 3=NoP, D, SD	NS	NS
INCUNDER (Incr. understand of why certain procs work well)	3/1=SA, 2=Ag, 3=Nop, Dis	NS	NS
NEWTECH (Incorp. new techniques)	3/1=SA, 2=Ag, 3=Nop, Dis	NS	NS
Amount of Coaching			
TRAINER (Ave obs by trainer/yr)	3/0, 1 &)=2	NS	NS
PRINPAL (Ave obs by princ./yr)	5/0, 1, 2, 3 &)=4	NS	NS
OFFCOACH (Ave obs by P+T/yr)	5/0, 1, 2, 3 &)=4	F_I=1.9 (P=.07)	F_I=2.0 (P=.05)
COACHYN (Any coaching?)	2/0=none, 1=some	NS	NS
Perceived Quality of Coaching Scales			
SCALE1 (Coach "focussed")	4/1=SA, 1.5, 2=Ag, 2.5=less positive	NS	NS
SCALE2 (Coach "analyzed")	4/same as SCALE1	NS	NS
SCALE3 (Coach "moderate")	4/same as SCALE1	NS	NS
SCALE4 (Coach "equivocation)	4/1=1-1.5, 2=2-2.5 3=3-3.5 & 4=4-5	NS	F=2.65 (P=.06)
Why Took PET Training			
MOTIVE (Reason for training)	3/1=vol, 2=Prin, 3=Other	NS	NS

Figure 1


BEDS FILE
Grade 1-4
Self-contained code
$|E_{MIN} - E_{MAX}| \leq 2$
L + M codes
$\bar{E} \geq 10$
NO
YES
n = 6049
SIRP Achievement Data File
Aggregation Based on Classroom Header
NO
n ≥ 10
YES
n = 8491
NO
COMPUTER MATCH
PET FILE
NO
SAME SCHOOL AND GRADE
YES
n = 5608
NO PET n = 4002
COMPUTER MATCH
PET ONLY DELETED
TEST COORDINATOR INFO
NOT OK
OK
n = 3483
MATCHED PET n = 1606
TEST COORDINATOR INFO
OK
NOT OK
n = 1305
VISUAL EDIT
n = 1279
CLASS SIZE DIFFERENCE ≥ 5
NO
YES
n = 1241
PET 2 n = 26
PET 1 n = 248
PET 0 n = 637
OTHER n = 330
VISUAL EDIT
OK
NOT OK
n = 3477
CLASS SIZE DIFFERENCE ≥ 5
NO
YES
n = 3378

Appendix

Matching Rules

BEDS to Achievement Data Matches

1. Last names must match except for obvious miscodings and/or name reversals
   Examples: LONG = LQNG
   ANDERSON JACQUELINE = JACQU LINE BNDERSON

2. Blank first names and/or initials are allowed since school and grade are controlled

3. Inverted first names will be allowed but they may not be contradictory except for nicknames obviously associated with a given first name.
   Examples: Becky = Rebecca
   Mary X Martha

BEDS/Achievement Data to PET Data Matches

1. Inverted first names will be allowed but they may not be contradictory except for nicknames obviously associated with a given first name.
   Examples: Becky = Rebecca
   Mary X Martha

2. For uncommon surnames in the three-way match (including the PET file with only district codes) inversions will be allowed and discrepant initials or blanks will be permitted.

3. For common surnames in the three-way match inversions will be allowed in any name. One difference of middle initial or associated name is the maximum allowable discrepancy.
   Examples: Keep permutations of the following:
   PHILLIPS MARY A = PHILLIPS MARY B =
   PHILLIPS A M = PHILLIPS ANITA M

   Delete the following:
   PHILLIPS MARY A = PHILLIPS MARY B = PHILLIPS MARY C

4. Where blank initials have been permitted in the BEDS to Achievement Data match, the PET file name may be considered to match if the maximum discrepancy is one initial or associated name.